Variation in performance on common content items at UK medical schools

Hope et al. BMC Medical Education (2021) 21:323
https://doi.org/10.1186/s12909-021-02761-1

BMC Medical Education

RESEARCH Open Access

Variation in performance on common content items at UK medical schools

David Hope[1*], David Kluth[1], Matthew Homer[2], Avril Dewar[1], Richard Fuller[3] and Helen Cameron[4]

Abstract

Background: Due to differing assessment systems across UK medical schools, making meaningful cross-school comparisons on undergraduate students' performance in knowledge tests is difficult. Ahead of the introduction of a national licensing assessment in the UK, we evaluate schools' performances on a shared pool of "common content" knowledge test items to compare candidates at different schools and evaluate whether they would pass under different standard setting regimes. Such information can then help develop a cross-school consensus on standard setting shared content.

Methods: We undertook a cross-sectional study in the academic sessions 2016-17 and 2017-18. Sixty "best of five" multiple choice 'common content' items were delivered each year, with five used in both years. In 2016-17 30 (of 31 eligible) medical schools undertook a mean of 52.6 items with 7,177 participants. In 2017-18 the same 30 medical schools undertook a mean of 52.8 items with 7,165 participants, creating a full sample of 14,342 medical students sitting common content prior to graduation. Using mean scores, we compared performance across items and carried out a "like-for-like" comparison of schools who used the same set of items then modelled the impact of different passing standards on these schools.

Results: Schools varied substantially on candidate total score. Schools differed in their performance with large (Cohen's *d* around 1) effects. A passing standard that would see 5 % of candidates at high scoring schools fail left low-scoring schools with fail rates of up to 40 %, whereas a passing standard that would see 5 % of candidates at low scoring schools fail would see virtually no candidates from high scoring schools fail.

Conclusions: Candidates at different schools exhibited significant differences in scores in two separate sittings. Performance varied by enough that standards that produce realistic fail rates in one medical school may produce substantially different pass rates in other medical schools – despite identical content and the candidates being governed by the same regulator. Regardless of which hypothetical standards are "correct" as judged by experts, large institutional differences in pass rates must be explored and understood by medical educators before shared standards are applied. The study results can assist cross-school groups in developing a consensus on standard setting future licensing assessment.

* Correspondence: david.hope@ed.ac.uk
[1]Medical Education Unit, Edinburgh Medical School, The Chancellor's Building, College of Medicine and Veterinary Medicine, The University of Edinburgh, 49 Little France Crescent, EH16 4SB Edinburgh, United Kingdom
Full list of author information is available at the end of the article

Introduction

Assessment in medical education should ensure doctors are competent, safe practitioners [1, 2]. Typically, candidates approaching registration must sit an "exit" assessment to confirm suitability to work as a doctor [3]. The defensibility of such assessments is of great importance in maintaining the quality of medical education and ensuring patient safety.

Evaluating such assessments can be difficult. In almost all regulatory environments doctors graduate from different institutions. Therefore, a range of institutional contexts, curricula, admissions policies, and resources produce doctors who are nominally equivalent, but differ in experiences [4]. Regulators seek to ensure equivalence across institutions by monitoring and enforcing a shared set of values and requirements [5].

As the content, structure, and weighting of exit assessments vary, direct comparisons across institutions are very difficult to carry out. Several partial solutions have been tested. One approach is to compare candidates on later – usually postgraduate – assessment which can act as a comparative measure. Research has shown that graduates of different medical schools exhibit large differences in performance on postgraduate assessments [6]. Relatedly, evidence has suggested that the performance of individual medical students and doctors exhibits at least moderate stability over time [7, 8] which suggests the variety of candidates applying to medical schools, or their experiences at medical schools, may create meaningful differences between cohorts upon graduation. Performance on (postgraduate) assessment predicts not just technical skill, but professionalism, including the likelihood of being sanctioned while working as a doctor [9, 10]. However, only a small proportion of doctors experience formal sanctioning, and written assessment is only one part of the wider process of professional evaluation.

Collectively, the research on postgraduate performance necessarily contains limitations. Postgraduate attainment can only measure capabilities some years after doctors begin work and cannot confidently identify the source of such differences. Postgraduate assessments are often specialised and sat by only a subset of doctors, and candidates who exit the profession soon after graduation will never sit them.

An alternative method lies in the use of "common content." Here, a group of institutions pool resources and share assessment content across institutions. So, a group of medical schools may share stations in a clinical examination, or multiple-choice questions (MCQs) in a written examination, with the remaining content set locally and independently. By evaluating both the approach to standard setting and the attainment of different cohorts, it is possible to get a better sense of how variable institutions are, within a single regulatory framework. Research on common content has suggested that different medical schools set very different standards for identical content. Research on MCQ-type written assessment has shown significant differences in medical school standard setting with typically medium effects, with the attendant risk that candidates who passed at institutions with lenient standards would have failed – and potentially not graduated – at institutions with more stringent standards for the same content [11]. A follow-up exploration of standard setting at some of the same schools described institutional, individual, and group factors combining to create highly unique standard setting procedures despite using the same content at all institutions [12].

Research on "common content" clinical examination stations have found similar problems, with standards for the same station varying by up to 13 % between the most lenient and stringent school [3]. Evidence on attainment, rather than standards, remains very sparse but some research on clinical examinations showed medical school cohorts scoring significantly differently on common content stations, in a pool of four medical schools [13].

This is extremely important as it suggests that, even if the content tested in different medical schools is equivalent, the local variability of standards may lead to candidates passing in some environments when they would have failed in others. Indeed, research has suggested that across many measures – content, type, duration, and standard setting – medical schools have a widely varying range of approaches [6, 14]. The fear that monitoring systems do not ensure comparability across institutions has led to recommendations for a knowledge test "licensing" assessment which acts as a single point of measurement for all candidates, alongside a locally designed and delivered clinical performance test, all within a complex regulatory framework [15]. The utility of this proposal remains contested. To some it represents the advance of a test-centric culture where learning is devalued [16] and educational diversity reduced [17]. To others, there are potentially significant benefits to patient safety by harmonising standards [15, 18]. This is an especially challenging area as stakeholders may prioritise different issues: while medical educators may see significant value in a broad range of experiences and curriculum designs, patients and regulators may prioritise high confidence in minimum safety standards. Such licensing assessments may identify genuine differences in attainment between schools, with implications for local standard setting, safety monitoring, and passing rates. Stakeholders might regard a licensing assessment as especially desirable if differences in passing rates are considered to reflect genuine differences in competence. Alternatively as, new doctors currently appear to

integrate well into the workplace when adequately supported, [19, 20] a licensing assessment could be an expensive and unnecessary addition.

The practical and theoretical challenges of implementing any multi-site assessment are significant. In the Netherlands, a progress test delivered across institutions has led to a more effective use of resources and enabled cross-school research, but also disagreements over item quality and logistical difficulties in organising the new assessment [21, 22]. In the United States, students have responded to the United States Medical Licensing Examination (USMLE) Step 1 with a range of effective self-directed learning behaviours to maximise the likelihood of passing [23]. However, the focus on the candidate's USMLE score has led authors to claim other aspects of performance – including achievements during medical school – have been under-valued, which has in turn led to reporting changes whereby only the candidate's pass/fail status is reported [24]. Such research demonstrates that cross-school assessment inevitably has serious implications for curriculum design and student learning even in areas which the assessment does not directly assess.

Despite the potentially significant impacts of a new licensing assessment on passing rates at medical schools, little is known about how such assessment might influence standard setting and pass rates. As a first step, medical educators at all affected schools should be aware of the relative performance of their students and the potential impact of different standard setting regimes, which can in turn help develop a consensus on how to standard set national licensing assessment in a way that recognises educational diversity while also ensuring patient safety.

To develop better evidence in this area, we used "common content" MCQs developed by the Medical Schools Council Assessment Alliance (MSCAA[11]) to compare candidates at 30 medical schools, evaluate performance differences across common content, and estimate the impact of different standards on pass rates ahead of the implementation of a licensing assessment in the United Kingdom.

Methods

Context and study design

In the United Kingdom, medical schools are regulated by the General Medical Council (GMC). The GMC sets the standards for undergraduate medical education and defines a series of high-level outcomes which medical schools must meet. [5] UK medical school programmes are typically five or six years long and begin with an introduction to the fundamentals of medicine, anatomy, healthcare in society, and collaborative working. In later years, students rotate through clinical attachments in which they develop knowledge and practical skills in a clinical environment. Before graduation, they sit both written and practical (i.e., clinical) assessment, which must be completed to a satisfactory level before graduating as a doctor. The quantity of assessment at UK medical school varies, with the amount of assessment (in minutes) differing for written ($M = 2{,}000$, $SD = 600$) and practical assessment ($M = 500$, $SD = 200$). [4] All medical schools in this sample set a locally developed written and clinical examination as part of their final assessment.

We undertook a cross-sectional study in academic sessions 2016-17 and 2017-18. The MSCAA organised 60 core items for participating schools in 2016-17 and 60 in 2017-18, with five used in both years. These were all "single best answer" multiple choice questions with one correct option and four distractors.

The items were curated by the MSCAA Final Clinical Review Group, which is made up of clinicians with expertise in both medical education and the relevant topic areas. Membership in the group rotates so that, over time, all UK medical schools contribute to this group. The common content items were designed to represent the full range of content areas (specialties and learning outcomes) regarded as "core" for a new UK graduate. Items were blueprinted against both GMC Outcomes for Graduates (e.g. "immediate care in medical emergencies") and content areas (e.g. "respiratory"). So, a question might be blueprinted to test a candidate's knowledge of respiratory medicine when providing immediate care in a medical emergency. In total 23 content areas were used with an average of 2.6 items per area.

Medical school representatives with expertise in standard setting were invited to comment on the suitability of items and the final set of common content items was designed to be maximally relevant for new doctors. Schools were not obliged to use all items. Items were embedded in each school's final written assessment [11, 12].

Participants

All UK medical schools were offered the opportunity to participate in the common content project. All items were delivered within an exit examination sat near the end of medical school. In 2016-17 30 medical schools undertook a mean of 52.6 common content items, with a total candidate number of 7,177. In 2017-18 30 medical schools undertook a mean of 52.8 common content items, with a total candidate number of 7,165, making for 14,342 sittings evaluated within this study. Full details can be found in Table 1. Medical schools had complete control over how many items to use and could use any combination of items.

Table 1 School performance.

2016 - 17							2017-18						
ID	N	Items used	1 SD below	1 SD above	2 SD below	2 SD above	ID	N	Items used	1 SD below	1 SD above	2 SD below	2 SD above
1617_x01	344	56	2 (3.57%)	11 (19.64%)	1 (1.79%)	0 (0%)	1718_x01	248	59	5 (8.47%)	5 (8.47%)	1 (1.69%)	0 (0%)
1617_x02	238	58	10 (17.24%)	11 (18.97%)	2 (3.45%)	1 (1.72%)	1718_x02	132	60	2 (3.33%)	18 (30%)	0 (0%)	2 (3.33%)
1617_x03	125	56	10 (17.86%)	5 (8.93%)	0 (0%)	2 (3.57%)	1718_x03	197	60	17 (28.33%)	2 (3.33%)	6 (10%)	1 (1.67%)
1617_x04	247	41	7 (17.07%)	1 (2.44%)	1 (2.44%)	0 (0%)	1718_x04	165	37	1 (2.7%)	27 (72.97%)	0 (0%)	6 (16.22%)
1617_x05	282	60	6 (10%)	3 (5%)	1 (1.67%)	0 (0%)	1718_x05	222	60	6 (10%)	14 (23.33%)	0 (0%)	0 (0%)
1617_x06	429	50	1 (2%)	8 (16%)	0 (0%)	1 (2%)	1718_x06	321	58	11 (18.97%)	5 (8.62%)	1 (1.72%)	0 (0%)
1617_x07	67	60	16 (26.67%)	4 (6.67%)	5 (8.33%)	1 (1.67%)	1718_x07	274	57	31 (54.39%)	1 (1.75%)	5 (8.77%)	0 (0%)
1617_x08	233	55	1 (1.82%)	22 (40%)	0 (0%)	3 (5.45%)	1718_x08	70	60	19 (31.67%)	7 (11.67%)	5 (8.33%)	0 (0%)
1617_x09	275	45	15 (33.33%)	2 (4.44%)	4 (8.89%)	0 (0%)	1718_x09	330	53	3 (5.66%)	11 (20.75%)	1 (1.89%)	0 (0%)
1617_x10	167	60	29 (48.33%)	1 (1.67%)	9 (15%)	0 (0%)	1718_x10	150	60	1 (1.67%)	11 (18.33%)	0 (0%)	1 (1.67%)
1617_x11	250	26	6 (23.08%)	1 (3.85%)	1 (3.85%)	1 (3.85%)	1718_x11	355	60	2 (3.33%)	34 (56.67%)	1 (1.67%)	6 (10%)
1617_x12	120	60	11 (18.33%)	8 (13.33%)	0 (0%)	1 (1.67%)	1718_x12	189	59	3 (5.08%)	16 (27.12%)	0 (0%)	0 (0%)
1617_x13	397	59	5 (8.47%)	9 (15.25%)	0 (0%)	2 (3.39%)	1718_x13	216	60	4 (6.67%)	3 (5%)	1 (1.67%)	0 (0%)
1617_x14	166	56	1 (1.79%)	20 (35.71%)	0 (0%)	0 (0%)	1718_x14	268	25	2 (8%)	2 (8%)	0 (0%)	2 (8%)
1617_x15	183	58	13 (22.41%)	7 (12.07%)	0 (0%)	1 (1.72%)	1718_x15	467	44	11 (25%)	1 (2.27%)	0 (0%)	0 (0%)
1617_x16	253	60	9 (15%)	2 (3.33%)	0 (0%)	0 (0%)	1718_x16	286	51	9 (17.65%)	5 (9.8%)	3 (5.88%)	3 (5.88%)
1617_x17	299	60	16 (26.67%)	2 (3.33%)	0 (0%)	0 (0%)	1718_x17	427	49	3 (6.12%)	7 (14.29%)	0 (0%)	1 (2.04%)
1617_x18	54	60	22 (36.67%)	5 (8.33%)	7 (11.67%)	0 (0%)	1718_x18	133	56	9 (16.07%)	10 (17.86%)	1 (1.79%)	2 (3.57%)
1617_x19	336	60	4 (6.67%)	15 (25%)	1 (1.67%)	0 (0%)	1718_x19	412	53	10 (18.87%)	2 (3.77%)	0 (0%)	0 (0%)
1617_x20	142	57	4 (7.02%)	10 (17.54%)	1 (1.75%)	0 (0%)	1718_x20	115	60	7 (11.67%)	10 (16.67%)	1 (1.67%)	0 (0%)
1617_x21	225	60	6 (10%)	6 (10%)	0 (0%)	0 (0%)	1718_x21	146	45	2 (4.44%)	5 (11.11%)	0 (0%)	0 (0%)
1617_x22	116	54	15 (27.78%)	3 (5.56%)	4 (7.41%)	0 (0%)	1718_x22	196	59	11 (18.64%)	3 (5.08%)	1 (1.69%)	0 (0%)
1617_x23	343	57	1 (1.75%)	10 (17.54%)	0 (0%)	1 (1.75%)	1718_x23	41	58	20 (34.48%)	5 (8.62%)	9 (15.52%)	0 (0%)
1617_x24	329	54	11 (20.37%)	4 (7.41%)	1 (1.85%)	2 (3.7%)	1718_x24	200	60	16 (26.67%)	1 (1.67%)	3 (5%)	0 (0%)
1617_x25	176	60	3 (5%)	22 (36.67%)	2 (3.33%)	0 (0%)	1718_x25	275	54	17 (31.48%)	2 (3.7%)	3 (5.56%)	0 (0%)
1617_x26	353	14	3 (21.43%)	1 (7.14%)	0 (0%)	2 (14.29%)	1718_x26	170	58	10 (17.24%)	7 (12.07%)	2 (3.45%)	1 (1.72%)
1617_x27	155	33	1 (3.03%)	23 (69.7%)	1 (3.03%)	10 (30.3%)	1718_x27	366	56	2 (3.57%)	5 (8.93%)	0 (0%)	1 (1.79%)
1617_x28	287	38	14 (36.84%)	0 (0%)	5 (13.16%)	0 (0%)	1718_x28	285	48	5 (10.42%)	10 (20.83%)	0 (0%)	3 (6.25%)
1617_x29	412	53	5 (9.43%)	1 (1.89%)	1 (1.89%)	1 (1.89%)	1718_x29	186	59	7 (11.86%)	16 (27.12%)	0 (0%)	0 (0%)
1617_x30	174	57	5 (8.77%)	9 (15.79%)	0 (0%)	0 (0%)	1718_x30	323	6	1 (16.67%)	2 (33.33%)	0 (0%)	2 (33.33%)

Note: School codes were different each year. "SD" = Standard Deviation

Ethics

Ethical approval was granted by the University of Edinburgh Medicine and Veterinary Medicine ethics committee. All participant details – both schools and candidates – were anonymised, and the research team had no access to deanonymized data.

Data collection

Following the completion of assessment, each school reported on the common content items to the MSCAA. This included notes from staff or candidates expressing concerns over item quality and a report of performance per candidate per question. The MSCAA then evaluated the psychometrics of the assessment using a combination of Classical Test Theory (CTT) and Rasch analysis to test whether items were of acceptable quality for analysis. Where a candidate failed to answer a question, this was coded as 0 (incorrect). An exploration of missing responses identified no pattern that would call into question the defensibility of any items or candidate response patterns.

Data analysis

As medical schools varied in the common content items they selected, making like-for-like comparisons was challenging. We utilised a two-part approach. In part 1, we compared means/facility scores, standard deviations, and discrimination indices for every item for every medical school that used the item. This allowed us to compare the homogeneity of medical schools in terms of both their average score and their variability. We sought to identify where (and how frequently) a given medical school significantly varied compared to other schools to see whether variability could be explained by small deviations across many items, or large deviations in a small number of items. This analysis was intended to be primarily descriptive, though we carried out a formal test of significance (via t-tests) for completeness.

In part 2, we selected a subsample of schools who had all sat a large proportion of the items. 13 schools sat the same 41 items in 2016-17, and 14 schools sat the same 48 items in 2017-18. Note that these numbers refer to the amount of overlapping, shared content. Some schools used more items, but as these were not shared by the entire subsample they were not included in this analysis.

To investigate whether this subset of items differed from the full, 60-item pool, we compared the blueprints of the used and unused item pools. We were unable to identify any pattern of differences in either outcomes or content areas. For example, items on "respiratory" were represented in both pools. These items might differ on

the combination of outcomes and content areas, or the specific aspects of respiratory knowledge being investigated, but the high-level outcomes and content areas were equally represented. This most likely represents the decision by the schools included in this subset to select a large, representative sample of items which covered most of the curriculum.

For these schools we carried out a like-for-like analysis of their within-year performance, tested whether performance of the top and bottom tertiles (representing "high scoring" and "low scoring" schools) differed significantly and modelled the impact of different passing standards. An a-priori power calculation showed that analyses used were able to detect small effect sizes at 80 % power [25]. School codes were not re-used, so the same code referred to a different school in each year.

Part 1 – item performance

We report here a Classical Test Theory (CTT) analysis of the data. While there are advantages to alternative methods – especially Rasch analysis [26] – the comparative simplicity and familiarity of CTT methods were considered desirable given the objective of maximising accessibility for the largest possible audience [27]. While we analysed the data using both a CTT and Rasch framework, only the CTT values are reported here.

For each item, we calculated the overall mean (or facility) score (between zero, indicating no candidate answered the item correctly, and one, indicating all candidates answered correctly), the Standard Deviation (SD) and the discrimination index (a measure of whether the item could discriminate between candidates who performed well or poorly on the assessment as a whole [28]). Facility and discrimination values did not differ significantly between the two study years, indicating the common content operated similarly in each year, and so we repeated the same analysis on each cohort. We calculated mean item facility ($M = 0.74$, $SD = 0.18$) and mean item discrimination on items ($M = 0.20$, $SD = 0.10$). We then calculated mean item performance (and associated SDs) for each school, per year. We then identified the proportion of items where the school was one or two SDs above the mean score, and one or two SDs below the mean score as a measure of the school's overall performance against all medical schools.

To further explore this, we compared the total number of items where the school scored two SDs below the mean. For the analysis, we compared the bottom and top tertiles and ran the analysis for each cohort. This gave a percentage measurement from zero (the school had no items 2 SDs below the mean) to 100 % (the school's cohort scored 2 SDs below the mean for every item). We calculated tertiles by the school's mean mark across all the items they used, and so compared the bottom tertile (the ten lowest performing medical schools on this assessment) against the top tertile (the ten highest performing medical schools).

The main goal of this was not to provide a precise comparison – because schools did not sit exactly the same items this was not possible – but to explore whether differences between schools could be explained by some schools exhibiting much higher rates of incorrect responses across a range of domains. Additionally, this relatively straightforward analysis can be reproduced by medical schools for internal evaluation and to address student queries, without requiring advanced statistical knowledge or significant researcher time. We chose to use 2 SDs as a cutoff as this generally indicated a notably lower score compared to the average school. The observed variance may then reflect differences in teaching approaches and curricula between medical schools, or genuine differences in student competence.

Part 2 – modelling standards

By comparing item usage across all schools, we identified schools which shared many items. We modelled the interaction of school numbers vs. item numbers: at one extreme it would be possible to compare all schools on a very small number of items, and at the other extreme a very small number of schools on all items. After modelling options, we were able to identify 13 schools from the 2016-17 cohort that had used the same 41 items, and a further 14 schools from the 2017-18 cohort that had used the same 48 items.

This gave us two samples of medical schools sitting identical content. For both years, Cronbach's alpha = 0.7, indicating an acceptable level of internal consistency for the two sets of items. We compared the bottom and top third of medical schools (rounded for uneven group sizes) in each sample on mean score. As in part 1, the sample size was adequate to test for small effects at 80 % power.

We then modelled the effect of different passing standards. We identified the pass score that would give a score as close as possible to a 5 % fail rate at (a) the four highest-scoring schools ("stringent") and (b) the four lowest scoring schools ("lenient"). This number was chosen to match the typical fail rate of the Prescribing Safety Assessment (PSA), an assessment sat by candidates across UK medical schools with similar features to future potential licensing assessments [29]. We then estimated the impact of imposing these passing standards on the medical schools. Medical schools received a copy of the results and were able to identify their own school (but not other schools).

Results

Part 1 – item performance

In 2016-17, schools in the lowest tertile (that is, their total score on the common items placed them in the lowest third when ranked by performance) had a number of items with facility scores two SD below the mean ($M = 7.81\,\%$, $SD = 4.4\,\%$) whereas the top tertile (upper third) had none, a significant difference ($t(9) = 5.61$, $p = .001$) with a large effect size ($d = 2.51$). This pattern was repeated in 2017-18 with the bottom tertile having some ($M = 6.62\,\%$, $SD = 4.19\,\%$) and the top tertile again having none, a significant difference ($t(9) = 5$, $p = .001$) with a large effect size ($d = 2.23$). This meant that for both years, schools in the bottom tertile reported significantly higher rates of items with facility scores two SD below the mean, indicating a different level of knowledge among those medical school students compared to the top tertile cohorts. This suggests that differences in scores may reflect differences in knowledge across a range of areas.

A full summary of the medical schools, the number of items they used, their scores relative to other medical schools, and their local sample size can be found in Table 1.

Part 2 – modelling standards

In 2016-17, comparing the bottom ($M = 0.76$, $SD = 0.1$) and top ($M = 0.85$ $SD = 0.08$) tertiles identified a statistically significant difference ($t(1570.1) = -20.82$, $p = .001$) with a large effect size ($d = 1.01$). This pattern was repeated in 2017-18 where comparing the bottom ($M = 0.68$, $SD = 0.1$) and top ($M = 0.78$, $SD = 0.09$) tertiles identified a statistically significant difference ($t(1562.5) = -20.5$, $p = .001$), again with a large effect size ($d = 1.02$).

The passing standards diverged with important practical consequences. In 2016-17, the stringent standard was 29.5 (71.95 %) and the lenient standard 24.5 (59.76 %), out of a total of 41. In 2017-18 the stringent standard was 29.73 (61.94 %) and the lenient standard 24 (50 %), out of a total of 48. Table 2 summarises the impact of these illustrative standards on pass rates: applying the most stringent standards to the lowest-scoring medical school would lead to a fail rate of 39.52 % in 2016-17 and 31.98 % in 2017-18. Conversely, applying the lenient standard would lead to one medical school in 2016-17 and four in 2017-18 having no failing candidates at all.

Discussion

This paper explores the use of "common content" items shared across UK medical schools, embedded in the knowledge test components of high-stakes, graduating level assessment. We show that candidates from different medical schools exhibit significant differences in scores on common content, and that these differences are partly generalisable – with schools differing across many domains. Importantly, a like-for-like comparison

Table 2 Like-for-like comparison across schools.

2016–17							2017-18						
School	Mean score	SD score	Tertile	95 % pass score	Stringent % failing	Lenient % failing	School	Mean score	SD score	Tertile	95 % pass score	Stringent % failing	Lenient % failing
x08	0.87	0.07	3	30	2.58 %	0 %	x11	0.82	0.08	3	32	0.85 %	0 %
x19	0.85	0.07	3	30	2.98 %	0.30 %	x12	0.76	0.09	3	30	4.76 %	0 %
x25	0.84	0.08	3	29	6.25 %	0.57 %	x10	0.76	0.09	3	29	5.33 %	0 %
x12	0.82	0.09	3	27	14.17 %	1.67 %	x05	0.75	0.1	3	28	9.91 %	0.90 %
x05	0.81	0.09	2	28	12.77 %	1.42 %	x02	0.75	0.09	2	29	6.82 %	0 %
x15	0.81	0.09	2	26	15.30 %	0.55 %	x29	0.75	0.09	2	28	8.60 %	0.54 %
x21	0.8	0.08	2	27	15.56 %	2.67 %	x28	0.72	0.1	2	27	15.09 %	0.35 %
x16	0.8	0.09	2	26	18.58 %	2.77 %	x13	0.72	0.1	2	27	17.13 %	1.85 %
x04	0.78	0.09	1	25	25.91 %	4.45 %	x01	0.72	0.1	2	26	15.32 %	0.40 %
x07	0.77	0.09	1	25	26.87 %	2.99 %	x20	0.72	0.09	1	28	13.04 %	1.74 %
x18	0.76	0.07	1	26	22.22 %	1.85 %	x22	0.68	0.11	1	25	27.04 %	2.55 %
x17	0.76	0.09	1	24	32.44 %	5.35 %	x24	0.67	0.11	1	23	24.50 %	7.50 %
x10	0.74	0.11	1	22	39.52 %	10.78 %	x08	0.67	0.09	1	25	30.00 %	2.86 %
							x03	0.67	0.11	1	23	31.98 %	5.58 %

Note: school codes were different each year. The "stringent" standard set the pass score as close as possible to yield a 5 % fail rate for the highest scoring medical schools. The "lenient" standard set the pass score as close as possible to yield a 5 % fail rate for the lowest scoring medical schools. For the columns "stringent" and "lenient" the values refer to the percentage of candidates at the medical school who have failed the assessment under the stringent/lenient criteria. Tertile 1 contains the lowest scoring schools

shows scores vary by enough that standard setting approaches that produce realistic fail rates – that is, fail rates that match those reported in similar assessments and for medical schools [29, 30] – may produce substantially different fail rates despite identical content and candidates being governed by the same regulatory environment. It is important for all medical educators – including those responsible for clinical teaching – to be aware of such trends and to contribute to ongoing discussions on how to reach a consensus on standard setting for national licensing assessment. Even if the standards here are taken as illustrative only, the observed variation in hypothetical passing rates emphasises the need for medical educators to agree whether standards should be uniformly applied, or locally-determined – as either approach will have substantial practical implications for any cross-institutional assessment. Within this discussion it will be important to reach a consensus on the minimally acceptable standard among all stakeholders – and determine whether the current approaches to training new doctors [19, 20] will be assisted by a licensing assessment. However, it is possible that the observed variation here reflects genuine differences in performance by schools in the top and bottom tertiles. If so, this could support the argument for the application of a national 'minimally acceptable' standard, albeit with complex consequences for schools at the extremes of performance, as the paper will next explore.

These findings extend and support previous research. They suggest that differences found in postgraduate attainment [6] may be partly attributed to differences in undergraduate medical education or attainment. The limited previous evidence of attainment variation on common content has been reinforced [13]. The emerging consensus that standard setting is a highly localised and subjective process influenced by contextual factors including local curricular differences [11, 12] offers insights into the attainment differences found here. Schools may be emphasising different areas and levels of knowledge, which then leads to significant differences on a shared assessment.

The evidence suggests that a common set of passing standards would impose high (or low) pass rates on some schools. That this is not happening currently could be explained by standard setters being heavily influenced by the performance of their local students rather than applying an arguably more objective national standard. Alternatively, it could be that material outside the common content is unique – implying less equivalence across schools. Differences in common content scores between schools may be due to differences in cohort ability, or variations in the format and emphasis of assessment at each institution.

If medical schools have divergent standards due to "localisation," significant disruption may occur if a single national standard is imposed. This may have substantial effects on passing rates and may disrupt workforce supply or affect stakeholder confidence in the exit assessment unless all stakeholders can work together to develop a sufficiently flexible approach that is acceptable to everyone.

This work shows that a shared regulatory environment alone does not necessarily develop homogeneity of performance, though it may have set an effective minimum standard if the standards of the lowest-performing medical school were found to be acceptable to all stakeholders. Importantly, however, given the known passing rates of UK medical schools, were such a "minimum standard" acceptable it would raise the concern that high-performing medical schools may be failing candidates who would be considered of passing quality by that minimum standard.

The extent to which educational diversity in content knowledge and topic specialisation is a desirable outcome [17] or a problem requiring regulation needs further discussion among educators and stakeholders. Either way, the experience of national assessment elsewhere suggests inevitable disruption during the implementation period [21–24].

The underlying ambiguity around current standard setting processes emphasises a challenge to medical education itself. If ongoing research on standard setting and empirical evidence suggests standard setting is not reproducible across time and contexts [11, 12] we must consider the impact on defensibility of assessments. We cannot judge from this work whether highly scoring medical schools are too stringent or whether lower scoring medical schools are too lenient or whether they are simply different in ways current regulatory processes fail to identify. It is extremely difficult to establish if there is a "correct" approach in a complex environment, and involvement of stakeholders throughout institutions affected by national licensing assessment is necessary.

Strengths and limitations

This study has several methodological strengths. The items have been reviewed and audited by experts then sat by many candidates across many institutions. This led to a high-quality dataset covering almost all candidates within a single regulatory environment. Our ability to compare schools on shared subsets of items allowed for a rigorous estimation of the impact of different standard setting regimes using empirical data. As such it serves as a plausible model for a future licensing assessment. The developers of the common content project (MSCAA) have a significant role in developing the national licensing assessment for UK medical schools, and

items similar to those selected in this study are likely to be used in the licensing assessment itself, adding further rigour to the work described. Importantly, we have opted for a widely understood, simple analytical approach via Classical Test Theory to make the results accessible to the largest possible audience of medical educators, policy makers and other stakeholders.

Despite this, there were limitations. The pool of items is smaller than would be expected in a full-sized examination, and candidates also sat locally developed items which could not be included in this analysis. Some schools used relatively few common content items and the mechanism by which schools select or reject items – or how they are integrated into wider assessment and teaching – remains underexplored. This study uses common content instead of a licensing assessment, and so a complete licensing assessment might exhibit different patterns of results. The comparisons made in part 1, while useful, were based on different items representing different content domains and may have varied in difficulty level. The study did not incorporate admissions data, and so cannot determine the extent to which cohort differences in early academic performance explain the variance in common content scores. Finally, while the accessibility of the work is a positive, more advanced methods such as Rasch inevitably offer additional analytic tools not employed in this analysis [27]. However, it should be noted that the Rasch model of this dataset did not contradict any of the findings set out here.

Future research

Future research should explore the stability of these trends and expand the availability of common content material to better compare medical schools. It is important to identify the mechanisms behind these differences (for which controlling for admissions or other assessment scores is especially important), and to ensure that a broad range of medical schools across the spectrum of performance are involved in standard setting any proposed licensing assessment. More generally, the subjectivity of standard setting methods suggests we must more thoroughly explore the link between performance at medical school and performance in the workplace – to see how graduates of different ability levels perform in work. Doing so will help ensure undergraduate medical education are appropriate to the role(s) candidates are trained for.

Throughout this paper we have noted the tension between the promotion of educational diversity (often prized by medical educators) and the need to ensure rigorous minimum safety standards. As part of the development of licensing assessments it would be beneficial for researchers to consult widely with patients to ensure licensing assessments can best meet public needs. Such work could include how best to manage educational diversity and how to approach cross-institutional differences in attainment. More research on this important policy area would be very useful.

Conclusions

This study has highlighted differences in performance across UK medical schools. It is essential all stakeholders work together to better understand these differences and determine the extent to which the differences reflect desirable educational diversity – or indicate a need for change.

Acknowledgements

Not applicable.

Authors' contributions

David Hope wrote the manuscript text and developed the main analysis. Avril Dewar and Matthew Homer contributed to the analysis and reporting of results. David Kluth, Richard Fuller, and Helen Cameron provided expertise on assessment, advice on analyses, and interpreting results. All authors contributed to the initial grant application that supported this work and the ethics application that allowed it to progress. All authors contributed to and revised the manuscript. The author(s) read and approved the final manuscript.

Funding

The Medical Schools Council Assessment Alliance funded this research.

Availability of data and materials

Due to the confidentiality and sensitivity of high-stakes assessment data, the datasets described in this study are not publicly available. If you wish for more information about the dataset or study, please contact David Hope (david.hope@ed.ac.uk).

Declarations

Ethics approval and consent to participate

All methods were carried out in accordance with relevant guidelines and regulations. Approval for the work was granted by the University of Edinburgh Medicine and Veterinary Medicine ethics committee. All participants provided informed consent to participate in the research via their institutions.

Consent for publication

Not applicable.

Competing interest

We declare that the authors have no competing interests as defined by BMC, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

Author details

[1]Medical Education Unit, Edinburgh Medical School, The Chancellor's Building, College of Medicine and Veterinary Medicine, The University of Edinburgh, 49 Little France Crescent, EH16 4SB Edinburgh, United Kingdom. [2]Leeds School of Medicine, Worsley Building, Leeds Institute of Medical Education, University of Leeds, LS2 9JT Leeds, UK. [3]School of Medicine, University of Liverpool, University of Liverpool, Cedar House, Ashton St, L69 3GE Liverpool, UK. [4]Aston Medical School, Aston University, 295 Aston Express Way, B4 7ET Birmingham, UK.

Received: 22 January 2021 Accepted: 17 May 2021
Published online: 05 June 2021

References

1. Cox M, Irby DM, Epstein RM. Assessment in medical education. N Engl J Med. 2007;356:387–96.

2. Norcini JJ, McKinley DW. Assessment methods in medical education. Teach Teach Educ. 2007;23:239–50.
3. Boursicot KA, Roberts TE, Pell G. Standard setting for clinical competence at graduation from medical school: a comparison of passing scores across five medical schools. Adv Health Sci Educ. 2006;11(2):173–83.
4. Devine OP, Harborne AC, McManus IC. Assessment at UK medical schools varies substantially in volume, type and intensity and correlates with postgraduate attainment. BMC Med Educ. 2015;15(1):146.
5. General Medical Council. Outcomes for Graduates. Manchester: General Medical Council; 2015.
6. McManus I, Elder AT, de Champlain A, Dacre JE, Mollon J, Chis L. Graduates of different UK medical schools show substantial differences in performance on MRCP(UK) Part 1, Part 2 and PACES examinations. BMC Med. 2008;6:5.
7. McManus I, Woolf K, Dacre J, Paice E, Dewberry C. The Academic Backbone: longitudinal continuities in educational achievement from secondary school and medical school to MRCP(UK) and the specialist register in UK medical students and doctors. BMC Med. 2013 Nov 14;11(1):242.
8. Hope D, Cameron H. Academic performance remains predictive over a five year medical degree. Innov Educ Teach Int. 2018;55(5):501–10.
9. Papadakis MA, Teherani A, Banach MA, Knettler TR, Rattner SL, Stern DT. Disciplinary action by medical boards and prior behavior in medical school. N Engl J Med. 2005;353:2673–82.
10. Wakeford R, Ludka K, Woolf K, McManus IC. Fitness to practise sanctions in UK doctors are predicted by poor performance at MRCGP and MRCP (UK) assessments: data linkage study. BMC Med. 2018;16(1):230.
11. Taylor CA, Gurnell M, Melville CR, Kluth DC, Johnson N, Wass V. Variation in passing standards for graduation-level knowledge items at UK medical schools. Med Educ. 2017;51(6):612–20.
12. Yeates P, Cope N, Luksaite E, Hassell A, Dikomitis L. Exploring differences in individual and group judgements in standard setting. Med Educ. 2019;53(9):941–52.
13. Chesser A, Cameron H, Evans P, Cleland J, Boursicot K, Mires G. Sources of variation in performance on a shared OSCE station across four UK medical schools. Med Educ. 2009;43:526–32.
14. MacDougall M. Variation in assessment and standard setting practices across UK undergraduate medicine and the need for a benchmark. Int J Med Educ. 2015;6:125–35.
15. Rimmer A. GMC will develop single exam for all medical graduates wishing to practise in UK. BMJ. 2014 Oct 1;349:g5896.
16. Van Der Vleuten CP. The assessment of professional competence: developments, research and practical implications. Adv Health Sci Educ. 1996;1:41–67.
17. Allawi L, Ali S, Hassan F, Sohrabi F. UKMLA: American dream or nightmare? Med Teach. 2016;38(3):320.
18. Archer J, Lynn N, Coombes L, Roberts M, Gale T, Bere SR de. The medical licensing examination debate. Regul Gov. 2017;11(3):315–22.
19. Burford B, Whittle V, Vance GH. The relationship between medical student learning opportunities and preparedness for practice: a questionnaire study. BMC Med Educ. 2014;14(1):223.
20. Blencowe NS, Van Hamel C, Bethune R, Aspinall R. 'From scared to prepared': targeted structured induction training during the transition from medical school to foundation doctor. Perspect Med Educ. 2015;4(2):90–2.
21. Schuwirth L, Bosman G, Henning RH, Rinkel R, Wenink ACG. Collaboration on progress testing in medical schools in the Netherlands. Med Teach. 2010 Jan 1;32(6):476–9.
22. Tio RA, Schutte B, Meiboom AA, Greidanus J, Dubois EA, Bremers AJA. The progress test of medicine: the Dutch experience. Perspect Med Educ. 2016 Feb;5(1):51–5.
23. Burk-Rafel J, Santen SA, Purkiss J. Study Behaviors and USMLE Step 1 Performance: Implications of a Student Self-Directed Parallel Curriculum. Acad Med. 2017 Nov;92(11S):S67.
24. Pershing S, Co JPT, Katznelson L. The New USMLE Step 1 Paradigm: An Opportunity to Cultivate Diversity of Excellence. Acad Med. 2020 Sep 1;95(9):1325–8.
25. Faul F, Erdfelder E, Lang A-G, Buchner A. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods. 2007;39(2):175–91.
26. Schumacker RE, Smith EV. A Rasch Perspective. Educ Psychol Meas. 2007;67:394–409.
27. Tavakol M, Dennick R. Psychometric evaluation of a knowledge based examination using Rasch analysis: An illustrative guide: AMEE Guide No. 72. Med Teach. 2013;35(1):e838–48.
28. Allen MJ, Yen WM. Introduction to measurement theory. Monterey, CA: Brooks/Cole; 1979.
29. Maxwell SRJ, Coleman JJ, Bollington L, Taylor C, Webb DJ. Prescribing Safety Assessment 2016: Delivery of a national prescribing assessment to 7343 UK final-year medical students. Br J Clin Pharmacol. 2017;83(10):2249–58.
30. Arulampalam W, Naylor RA, Smith JP. A hazard model of the probability of medical school drop-out in the UK. J R Stat Soc Ser A Stat Soc. 2004;167:157–78.

Publisher's Note